International Journal of Electronics,
Communication & Instrumentation Engineering
Research and Development (IJECIERD)
ISSN(P): 2249-684X; ISSN(E): 2249-7951
Vol. 4, Issue 2, Apr 2014, 25-32
© TJPRC Pvt. Ltd.

LINE SEGMENTATION OF DEVNAGRI HANDWRITTEN DOCUMENTS

PRAMOD S. MALGI & SHAILJA GAYAKWAD

Department of Electronics Engineering, Terna Engineering College, University of Mumbai, Nerul,
Navi Mumbai, Maharashtra, India

ABSTRACT

In this paper, we have discussed the new method for Line Segmentation of Handwritten Hindi text. The method is based on header line detection, base line detection and contour following technique. No preprocessing like skew correction, thinning or noise removal has been done on the data. The purpose of this paper is three fold. Firstly, we explained by experiments that this method is suitable for fluctuating lines or variable skew lines of text. Also we confirm that this method is invariant of non uniform skew between words in a line (non uniform text line skew). Secondly, the contour following after header line detection correctly separates some of the overlapped lines of text. Thirdly, this paper provides a brief review of text line segmentation techniques for handwritten text which can be very useful for the beginners who want to work on text line segmentation.

KEYWORDS: Text Line Segmentation, Projection, Contour Following, Header Line, Base Line Detection

1. INTRODUCTION

Handwritten Hindi text recognition is an important area of Optical Character Recognition (OCR). A good survey about OCR is given in [1]. Devanagari is the script for writing Hindi text. Hindi is the official language of India. There are a number of applications where Handwritten Hindi Recognition systems can be used. Segmentation is an important stage of OCR system because many errors in recognition system occur due to errors in segmentation. The presence of upper and lower modifiers in Hindi text makes the line segmentation even more complex. Due to different writing styles of the people the lines of handwritten documents are not straight. The main problem of segmentation is skew in a text. The skew may be uniform or non uniform in whole of the text or in a line. Another problem is small gap between the lines or overlapping of the text lines. The interline distance may vary due to writer movement. Due to above problems the text line segmentation methods cannot be directly applied on handwritten documents.

In the past few years, Handwritten Hindi text recognition has gained a lot of attention of pattern recognition researchers. The text line segmentation methods can be generally categorized into bottom-up and top-down. In the bottom up approach, the grouping methods merge neighboring connected components using simple rules on the geometric relationship between neighboring blocks.

The projection based methods are the top-down algorithms which are one of the most successful methods for machine printed text. The projection based methods are also successful for handwritten text where text lines are straight or easily separable. But due to different writing styles of the people, the text line segmentation is still very challenging. The main purpose of this paper is to provide a method for line segmentation of fluctuating lines or skew variable handwritten Hindi text. The method is based on header line and base line detection. But to the best 2. In Section 3 we have discussed the creation of database used for the experimental purposes. Section4 contains the characteristics of Hindi language.

Section 5 includes the discussion about the line segmentation method. Finally, Section 6 contains results and Section 7 contains discussions on the work.

2. LITERATURE REVIEW

A lot of research is done in the past on line segmentation of handwritten text. A wide variety of line segmentation methods for handwritten documents are reported in the literature. The various existing methods for line segmentation are categorized as projection based, Hough transform based, smearing, grouping, graph based, CTM (Cut text Minimum) approach, Block covering and linear programming.

In projection based method horizontal projections of the pixels are used to segment the text into lines. This method is suited for straight lines or easily separable lines only. This method is modified by some researchers using partial projection method [2]. The input text is divided into vertical stripes and horizontal projection of each stripe is considered for line segmentation. Tripathy and Pal [3], also used a projection based method in text line stripes and combined the results of adjacent stripes into complete text line. Stripe-wise horizontal histograms are then computed and the relationship of the peak–valley points of the histograms are used for line segmentation of Oriya text. Width of the stripe is calculated after analyzing height of water reservoirs.

These methods can tolerate the skews in a text but cannot work for segmentation of overlapped or connected lines of text. In Hough transform based methods, the gravity centers or minima points are used for line segmentation. In [4], the text line detection method for unconstrained handwritten documents based on Hough transform is used. This method is not suitable for variable skewed lines.

In smearing method, fuzzy run length is used for line segmentation. The fuzzy run length describes how far one can see when standing at a pixel along horizontal direction. In [5], the smearing method is used for text line segmentation. Run length smearing algorithm is used to segment individual text lines from document images. The threshold for RLSA is computed based on the height information of the text lines.

In grouping method, starting from the bottom the connected components of black pixels are grouped together to form the groups or alignments. In [6], a of my knowledge this is the first paper on text line segmentation of handwritten Hindi text. In next Sections, we have reviewed the literature in Section approach based on perceptual grouping of black pixels is used for text line detection. In graph based method, a graph of main strokes is build from the document image and search is made for the shortest spanning tree. In [7], the graph based method is used for line segmentation.

In CTM [8], the path between text lines to be separated is searched which minimizes the text line pixels cut during line segmentation. The projection is used for rough estimate of text line separations. Then horizontal path is followed at the rough estimated line separation point to cut the minimum of text, especially descender from the upper line and ascenders from the lower line. This method [8], shows accuracy up to 96%.

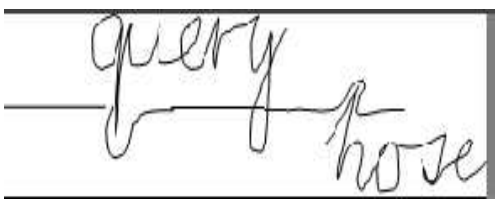

Figure 1: CTM Method [8]

In [9], the block covering technique is also used for segmentation of overlapped and multi-touching text lines of ancient Arabic documents. An optimal block covering is applied on tightly spaced documents. The large blocks generated by the block covering are then segmented by relying on the statistical analysis of block heights. The final labeling into text-lines is based on a block neighboring analysis. Pal and Dutta [10], also used the stripe based partial projection based method with water reservoir concept for line segmentation of Unconstrained Bangla Handwritten text. Water reservoirs are used to determine the height of the character. Lehal and Singh [11], presented an OCR for printed Gurumukhi script. They have also used horizontal projection profiles for line segmentation. The work on Handwritten Gurmukhi Script is still going. The techniques based on linear programming are also used for text line segmentation. The above methods are applied on printed or handwritten texts of Indian scripts like Bangla, Arabic, Gurmukhi and Oriya. The OCR's of other languages like Assamese, Tamil, Kannada, Gujarati and Malayalam are also developed by researchers and efforts are still going on to improve their recognition rates. In most of the OCR's, horizontal projection profiles are used for line segmentation. Pal and Chaudhuri [12], presented a complete OCR for printed Devanagari script. The text line is divided into three zones or stripes for segmentation and projection profile is used for segmentation. Among some latest work, Jindal *et. al*. [13-15] have worked on recognition of degraded various problems faced during recognition. The research on handwritten Devanagari text line segmentation is still going on. Devanagari is the script for writing Hindi language. Hindi is written from left to right and it also has half characters which makes the development of Handwritten Hindi OCR complex. Many OCR's for Indian Scripts are reported in the literature but very few researchers have worked on handwritten Hindi OCR. So our work is on segmentation of Handwritten Hindi text.

3. DATABASE

All experiments are conducted on database constructed by taking handwritten data from 15 writers and some data from different papers. Each writer was asked to write 5 to 15 lines of text. A healthy mix of people from various backgrounds was taken so as to make such a small database as close as possible to the real database. Total lines are 200 in number. No pre-processing like skew removal, noise removal is performed on the data.

4. CHARACTERISTICS OF HINDI/MARATHI LANGUAGE

Devanagari is the script for writing Hindi, Nepali, Marathi and Sanskrit languages. The alphabets of Devanagari script consists of 33 consonants and 14 vowels. It is written from left to right. There is no concept of lower or upper case in Hindi language.

In Hindi language, most of the characters have a horizontal line at the upper part. In Hindi language characters also have a half form which increases the language complexity for recognition. The half characters may touch with full characters to make the characters called conjuncts. In each conjunct character, the right part is a full consonant, and the left part is always a half consonant. When two or segmentation of handwritten Hindi text very complex.

5. LINE SEGMENTATION

We proposed a line segmentation method which is based on header line detection and base line detection. We have used two-stripe projection method for segmentation of lines. Header line is the most visible part of the text. Detection of header line is one the most challenging tasks in skew variable or fluctuating line text. Till now most of the researchers are detecting the header line by finding the row with maximum pixel density, but it cannot work for skew variable text. The method of text line segmentation proposed in this paper is extension of the work done by us in [16].

The algorithm for header line detection given by us in [16] cannot work if there is no row with maximum number pixels followed by 12 rows with pixels less than that row. So for header line detection, the condition in step 1 is modified in the algorithm proposed below.

We make following assumptions about the data:

- The minimum height of a consonant in a line is eight pixels. The average height of a line is between 20 to 40 pixels.
- The maximum height of a (consonant + lower modifier) is 25 pixels.
- The skew between two lines in a text is not more than the height of a consonant.

For finding the header line in skew variable text following procedure is adopted:

Step 1: Initially, rough estimate of the header lines in whole of the text are made by the formula

pcol(i)>15 & (pcol(i)>pcol(i:i+12)|flag2==1) &

pcol(i:i+8)>0 & pcol(i)>wdth(i)/7

where,

pcol(i): No. of pixel in row i.

wdth(i): width of line i i.e difference between last

pixel and first pixel position of line i.

flag2:

if pcol(i)<=2

More characters are combined to form a word, the horizontal lines touch each other and generate a header line called shirorekha. The vowels (modifiers) can be placed at the left, right (or both), top or bottom of the consonant. The vowels above the header line are called ascenders or upper modifiers and vowels below the consonants are called descenders or lower modifiers. Two consecutive lines touch or overlap each other due to these modifiers. This makes the

printed Gurmukhi script documents and addressedflag1=1;

hline=i;

end

if flag1==1& pcol(i)>15

lc=lc+1;

if lc>10

flag2=1;

end

end

hline and lc are the variables for line number and line count.

Step 2: After finding first header line, we skip 8 rows (equal to minimum height of consonant) to find the next header line.

Step 3: From (i+8)th row to (i+25)th row, we find the mth row with minimum of pixels.

Step 4: We skip the rows up to mth row and go to step 1 to find the next header line.

After finding the header lines, the most challenging task is to find the base line. For finding the base line following procedure is followed:

Step 1: Two consecutive rough header lines are taken.

Step 2: The line is divided into four equal parts (stripes).

Step 3: The rows with minimum of pixels are taken as base lines separately for each part.

Step 4: Then the lines are separated between header lines and base lines separately for each part.

Step 5: Then four separate lines are joined to get the actual text line.

This method gives good results for uniform and non uniform skewed lines. This method is also useful if the overlap is not within a strip but it between the strips. This method is not suitable for touching or overlapped lines within a strip. For overlapped lines we need modified this algorithm.

6. RESULTS

The accuracy of line segmentation depends upon the accuracy of header line and base line detection.

The header lines and base lines are accurately detected from the above database. The skewed text lines are also segmented accurately. The overlapped lines with complete ascenders (no broken parts) are also recognized correctly.

Following text message is used for line segmentation and the results are mentioned below.

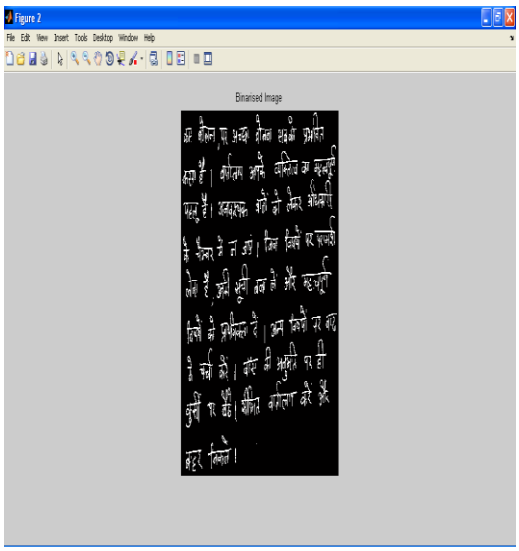

Figure 2: Sample of Hindi Text

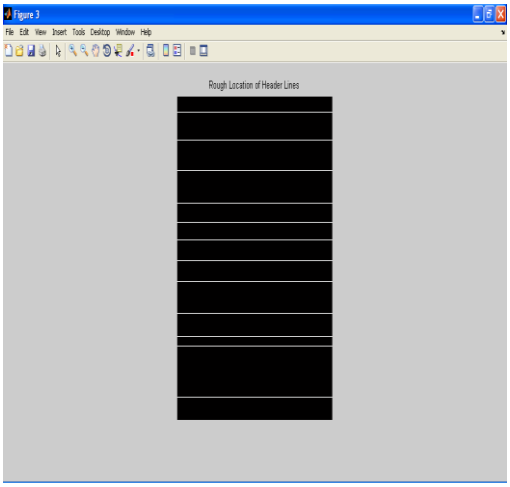

Figure 3: Header Line Detection

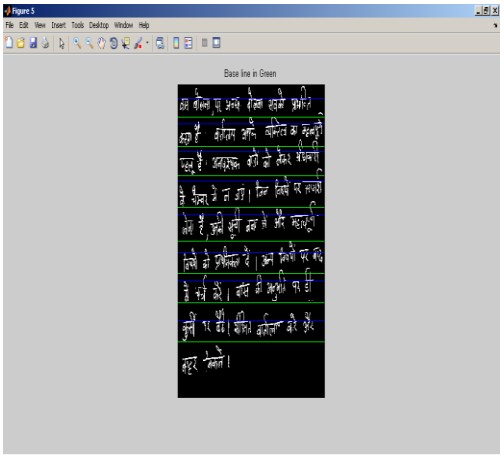

Figure 4: Header Line (Blue) & Base Line (Green)

Accuracy of segmentation

Total Lines	HeaderLines correctly Segmented	% Accuracy
9	7	78%

Total Lines	Base Lines correctly Segmented	% Accuracy
9	8	89%

Figure 5

7. DISCUSSIONS

From the above results of line segmentation it is clear that the header and base line detection method is very useful for line segmentation of variable skew or fluctuating lines of Handwritten Hindi text. The lines which have overlapping within a stripe are not correctly recognized by this method. For overlapping lines within a stripe, we tried to follow contour from the detected header lines. The lines which have broken parts in upper modifiers are not correctly recognized. The lines with thick parts in upper modifiers also not correctly recognized.

The study may be carried out on in future with following direction:

The above method can be applied on other languages like Bangla, Telugu etc. Some other technique may be tried for segmentation of overlapped or touching lines within a stripe.

8. REFERENCES

1. S. Mori, C. Y. Suen, and K. Yamamoto, “Historical review of OCR Research and development”, Proceedings of the IEEE, Vol. 80, No. 7, pp. 1029-1058, 1992.
2. A. Zahour, B. Taconet, P. Mercy, and S. Ramdane, “Arabic Hand-written Text-line Extraction”, Proceedings of the Sixth International. Conference on Document Analysis and Recognition, ICDAR, pp. 281–285, 2001.
3. N. Tripathy and U. Pal., “Handwriting Segmentation of unconstrained Oriya Text”, International Workshop on Frontiers in Handwriting Recognition, pp. 306–311, 2004.
4. G. Louloudis, B. Gatos, I. Pratikakis and K. Halatsis” A Block Based Hough Transform Mapping for Text Line Detection in Handwritten Documents”, Proceedings of the Tenth International Workshop on Frontiers in Handwriting Recognition, pp.515-520, 2006.
5. Y. Li, Y. Zheng, D. Doermann, and S. Jaeger, “A new algorithm for detecting text line in handwritten documents”, Proceedings of the Tenth International Workshop on Frontiers in Handwriting Recognition, pp. 35–40, 2006.
6. L. Likforman-Sulem and C. Faure, "Extracting text lines in handwritten documents by perceptual grouping", Advances in handwriting and drawing: a multidisciplinary approach, pp. 21-38, 1994.
7. I.S.I. Abuhaiba, S. Datta and M.J.J. Holt, "Line Extraction and Stroke Ordering of Text Pages", Proceedings of the Third International Conference on Document Analysis and Recognition, Canada, pp. 390- 393, 1995.
8. C. Weliwitage, A. L. Harvey and A. B. Jennings, “Handwritten Document Offline Text Line Segmentation”, Proceedings of Digital Imaging Computing: Techniques and Applications, pp. 184-187, 2005.
9. A. Zahour, B. Taconet, L. Likforman-Sulem and Wafa Boussellaa, “Overlapping and multi-touching textline segmentation by Block Covering analysis”, Pattern Analysis and Applications, Vol. 12, pp. 335-351, 2008.
10. U. Pal and S. Datta, “Segmentation of Bangla Unconstrained Handwritten Text”, Proceedings of the 7th International Conference, ICDAR, pp.1128-1132, 2003.
11. G. S. Lehal and Chandan Singh, “A Gurumukhi Script recognition system”, Proceeding of 15th International conference on Pattern recognition, Spain, Vol. 2, pp.557-560, 2000.
12. U. Pal and B. B. Chaudhuri, “Printed Devanagari Script OCR System”, Vivek, Vol. 10, pp.12-24, 1997.
13. M. K. Jindal, G. S. Lehal and R. K. Sharma, “On Segmentation of touching characters and overlapping lines in degraded printed Gurmukhi script”, International Journal of Image and Graphics (IJIG), World Scientific Publishing Company, Vol. 9, No. 3, pp. 321-353, 2009.

14. M. K. Jindal, R. K. Sharma and G. S. Lehal, "Segmentation ofHorizontally Overlapping Lines in Printed Indian Scripts", International Journal of Computational Intelligence Research (IJCIR), Research India Publications, vol. 3, No. 4, pp. 277-286, 2007.

15. M. K. Jindal, R. K. Sharma and G. S. Lehal, "Segmentation of Touching Characters in Upper Zone in printed Gurmukhi Script", Proceedings of the 2nd Bangalore Annual Compute Conference, Banglore, ACM, No. 9, 2009.

16. Naresh Kumar Garg, Lakhwinder Kaur and M. K. Jindal, "Segmentation of Handwritten Hindi Text", Accepted for Publication in International Journal of Futuristic Computer Applications (IJFCA), Vol. 1, 2010. 397